Bias and Fairness in Large Language Models: A Survey

Year: 2023 **|** *Citations: 858* **|** *Authors: Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow*

Abstract

Abstract Rapid advancements of large language models (LLMs) have enabled the processing, understanding, and generation of human-like text, with increasing integration into systems that touch our social sphere. Despite this success, these models can learn, perpetuate, and amplify harmful social biases. In this article, we present a comprehensive survey of bias evaluation and mitigation techniques for LLMs. We first consolidate, formalize, and expand notions of social bias and fairness in natural language processing, defining distinct facets of harm and introducing several desiderata to operationalize fairness for LLMs. We then unify the literature by proposing three intuitive taxonomies, two for bias evaluation, namely, metrics and datasets, and one for mitigation. Our first taxonomy of metrics for bias evaluation disambiguates the relationship between metrics and evaluation datasets, and organizes metrics by the different levels at which they operate in a model: embeddings, probabilities, and generated text. Our second taxonomy of datasets for bias evaluation categorizes datasets by their structure as counterfactual inputs or prompts, and identifies the targeted harms and social groups; we also release a consolidation of publicly available datasets for improved access. Our third taxonomy of techniques for bias mitigation classifies methods by their intervention during pre-processing, in-training, intra-processing, and post-processing, with granular subcategories that elucidate research trends. Finally, we identify open problems and challenges for future work. Synthesizing a wide range of recent research, we aim to provide a clear guide of the existing literature that empowers researchers and practitioners to better understand and prevent the propagation of bias in LLMs.